Comparing Variable Selection

Simulation Settings

The experiments simulate competing risks data from two primary frameworks: cause-specific proportional hazards models and a proportional sub-distribution hazards model. For all simulations, the sample size is fixed at $n = 400$, while the number of covariates, p, is varied between a low-dimensional case ($p = 120$) and a high-dimensional case ($p = 1000$).

Data is primarily generated under a cause-specific proportional hazards framework using a Weibull baseline hazard. The hazard function for cause j ($j = 1, 2$) for an individual i with covariates X_i is given by

$$\alpha_{ij}(t; X_i) = \lambda_{0j} \exp(\beta_j^T X_i)\gamma_j t^{\gamma_j - 1}$$

where λ_{0j} is the baseline hazard rate parameter and γ_j is the shape parameter for the Weibull distribution. In the provided simulations, the shape parameters are fixed at $\gamma_1 = \gamma_2 = 1.5$. The baseline hazard rates are set to $\lambda_{01} = 0.55$ and $\lambda_{02} = 0.35$. Censoring times are generated from an exponential distribution, calibrated to achieve approximately 25%. The simulations consider two events ($j = 1, 2$) with corresponding coefficient vectors β_1 and β_2, each of length p. A subset of these coefficients, corresponding to the k true variables, are non-zero. The non-zero coefficients are defined across five different settings:

- **Setting 1: Single effects on endpoint of interest.** Only the first 20 variables are associated with the cause-specific hazard of event 1, and no variables are associated with event 2. The coefficient vectors are

$$\beta_1 = (\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{p-k})^\top, \quad \beta_2 = (\underbrace{0, \ldots, 0}_{p})^\top$$

- **Setting 2: Single effects on both endpoints.** Two distinct sets of 10 variables each influence the two hazards separately.

$$\beta_1 = (\underbrace{1, \ldots, 1}_{k/4}, \underbrace{0, \ldots, 0}_{k/4}, \underbrace{1, \ldots, 1}_{k/4}, \underbrace{0, \ldots, 0}_{k/4}, \underbrace{0, \ldots, 0}_{p-k})^\top, \quad \beta_2 = (\underbrace{0, \ldots, 0}_{k/4}, \underbrace{1, \ldots, 1}_{k/4}, \underbrace{0, \ldots, 0}_{k/4}, \underbrace{1, \ldots, 1}_{k/4}, \underbrace{0, \ldots, 0}_{p-k})^\top$$

- **Setting 3: Opposing effects.**}A set of 20 variables has opposing effects on the two event types.

$$\beta_1 = (\underbrace{+0.5, -0.5, \ldots, +0.5, -0.5}_{k}, \underbrace{0, \ldots, 0}_{p-k})^\top, \quad \beta_2 = (\underbrace{-0.5, +0.5, \ldots, -0.5, +0.5}_{k}, \underbrace{0, \ldots, 0}_{p-k})^\top$$

- **Setting 4: Mixture of effects.** This scenario involves a mix of single and opposing effects as shown below.

$$\beta_1 = (\underbrace{1, \ldots, 1}_{k/4}, \underbrace{0.5, -0.5, \ldots, 0.5, -0.5}_{k/4}, \underbrace{1, \ldots, 1}_{k/4}, \underbrace{0, \ldots, 0}_{k/4}, \underbrace{0, \ldots, 0}_{p-k})^\top,$$

$$\beta_2 = (\underbrace{0,\dots,0}_{k/4}, \underbrace{-0.5, 0.5, \dots, -0.5, 0.5}_{k/4}, \underbrace{0,\dots,0}_{k/4}, \underbrace{1,\dots,1}_{k/4}, \underbrace{0,\dots,0}_{p-k})^\top$$

- **Setting 5: Non-proportional hazards.** This setting is simulated under the sub-distribution hazards framework, which violates the proportionality assumption for the cause-specific hazards. The coefficient structure is specified as:

$$\beta_1 = (\underbrace{1,\dots,1}_{k}, \underbrace{0,\dots,0}_{p-k})^\top, \quad \beta_2 = (\underbrace{-1,\dots,-1}_{k}, \underbrace{0,\dots,0}_{p-k})^\top$$

The covariates X are generated from a multivariate normal distribution with mean $\mu = 0$ and a specified correlation structure. For settings 1 and 5, an exchangeable correlation structure is used among the true positive variables with $\rho = 0.5$. For settings 2, 3, and 4, a block correlation structure with 4 blocks is used for the true positive variables. In all settings, the noise variables are generated with an exchangeable correlation of $\rho = 0.1$.

For setting 5, data is generated from a proportional sub-distribution hazards model as specified by Fine and Gray (1999).

Competing Algorithms

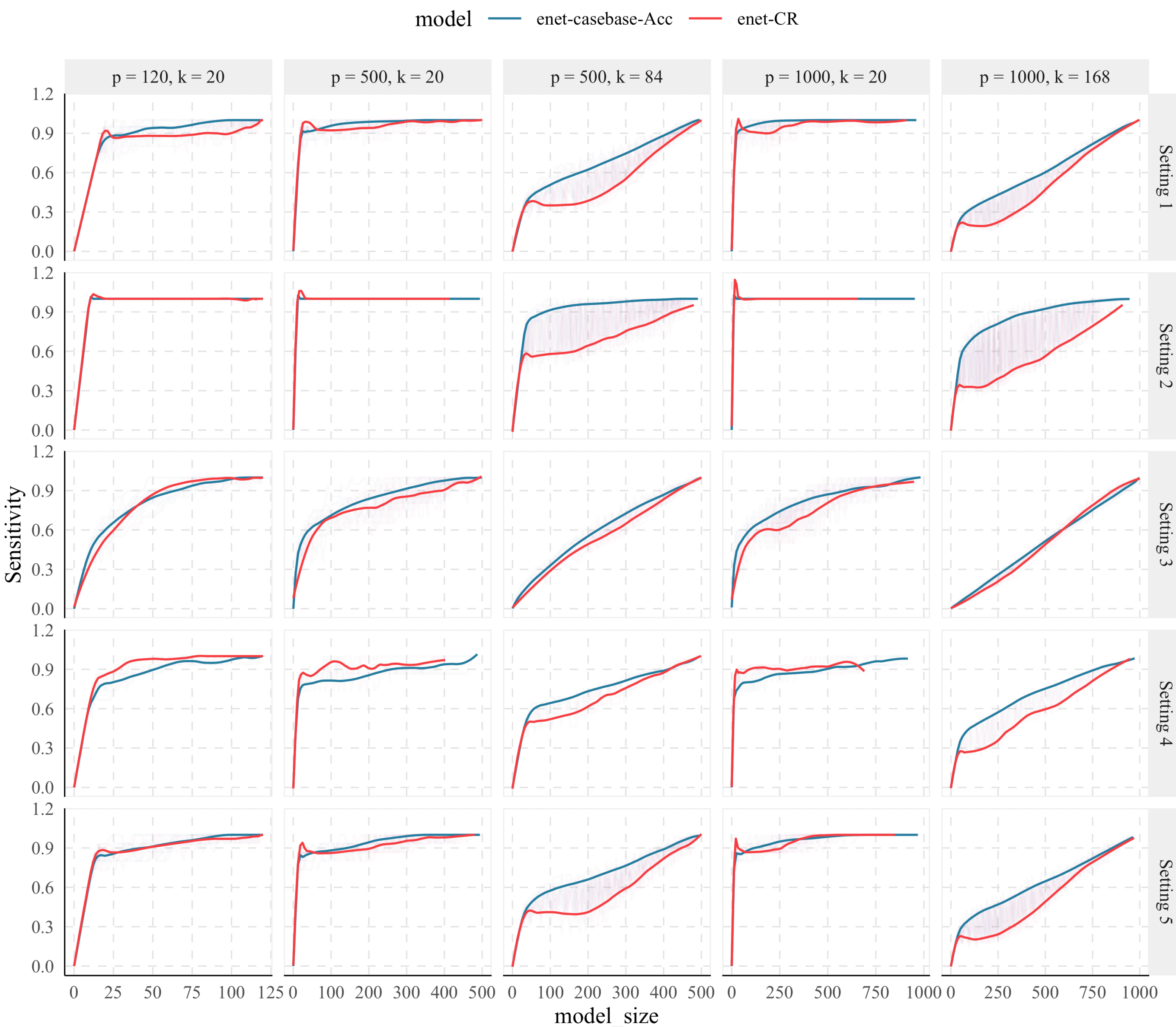
In this document, two algorithms are compared in terms of selection performance.

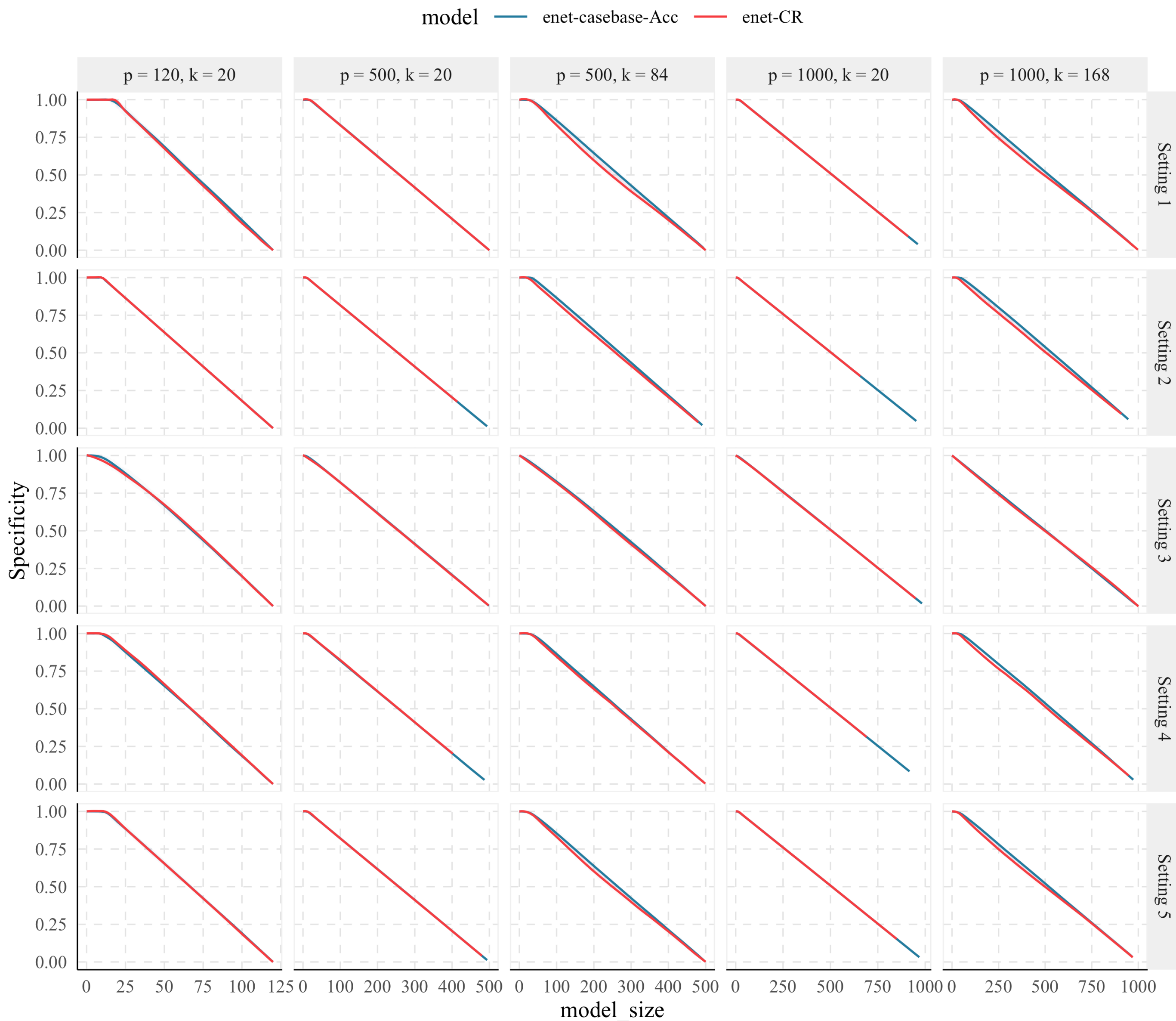
- **Elastic-net case-base (enet-casebase).** This method uses case-base sampling with an elastic-net penalization
- **Elastic-net Independent Cox (enet-iCR):** This approach fits separate elastic-net penalized Cox models for cause 1.

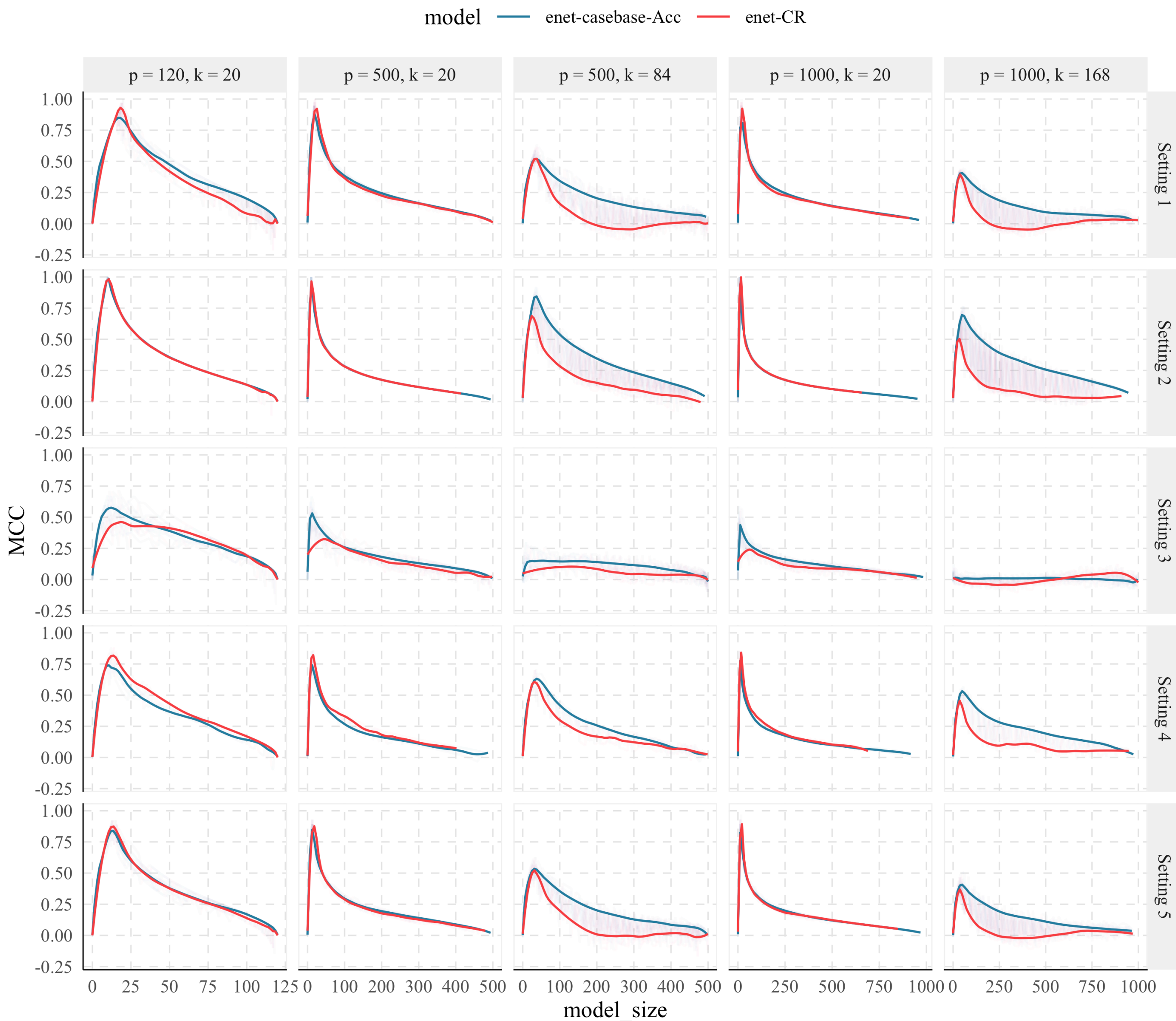
Results

For each simulation, the algorithms were fitted using the lambda values generated through cross-validation, although cross-validation was not actually performed in this case. The number of lambda values was set to 50. Consequently, the plots display smoothed lines for each simulation and setting. It's worth noting that the average results appear more scattered. Therefore, I opted for the smoothed lines to provide a clearer comparison.

1 Simulations




model
enet-casebase-Acc
enet-CR
p = 120, k = 20
p = 500, k = 20
p = 500, k = 84
p = 1000, k = 20
p = 1000, k = 168
Setting 1
Setting 2
Setting 3
Setting 4
Setting 5
Specificity
model_size



References

Fine, Jason P., and Robert J. Gray. 1999. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94 (446): 496–509. https://doi.org/10.2307/2670170.